Crime Analysis of Chicago

Shuai Wang, Weifeng Li
Department of Computer Science and Engineering
University of Bridgeport, Bridgeport, CT

Abstract

Security status in the place where we are live is one of the most important concerns for every individual. As most people may know, to evaluate the security status, criminal rate cannot be overlooked. Many factors can influence the occurrence of crimes including time, places, the construction of population, educational level, income level and so on. Nowadays, using modern technology such as Hadoop to process large volume of crime dataset, analyze the relationship between crime rate and the related factors and predict the happen of crime are getting more and more popular. In this project, using Hadoop ecosystem and related technologies, we choose crimes dataset of Chicago from 2001 to present to analyze the relationships between the occurrence of crimes and several key factors.

Dataset

Two data sources were chosen by us as the input of our Hadoop Map-Reduce program. One is the crime dataset in City of Chicago from 2001 to present. The total size of this dataset is 1.5G and is stored as a CSV format file. It reflects reported incidents of crime that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is extracted from the Chicago Police Department's Citizen Law Enforcement Analysis and Reporting system. In order to protect the privacy of crime victims, addresses are shown at the block level only and specific locations are not identified. The second dataset is the socioeconomic indicators in Chicago which reflects the economic status in different communities of Chicago. This dataset contains a selection of six socioeconomic indicators of public health significance and a "hardship index", by Chicago community area, for the years 2008 – 2012. Part of the most import attributes of the two dataset are shown as below.

ID	Date	Primary Type	Arrest	Domestic	Community	Year

Community Number	Community Name	Per Capital Income	Hard Index

Our goal is to process the datasets extracting the most important information from them and then draw the relationships between happening of crimes and these factors.

Problem Definition

To achieve our goal, we try to solve four problems in our project.

- First of all, we want to know the crime trend year by year from which we can know whether the security status in the city is getting better or worse.
- Second, what time is relatively more secure and which time period is more dangerous are also important for us.
- Besides, as we all know, different crime types have different level of harm, we are interested in the proportion of different types of crime.
- Finally, how the socioeconomic indexes affect the occurrence of crimes are our last concern.

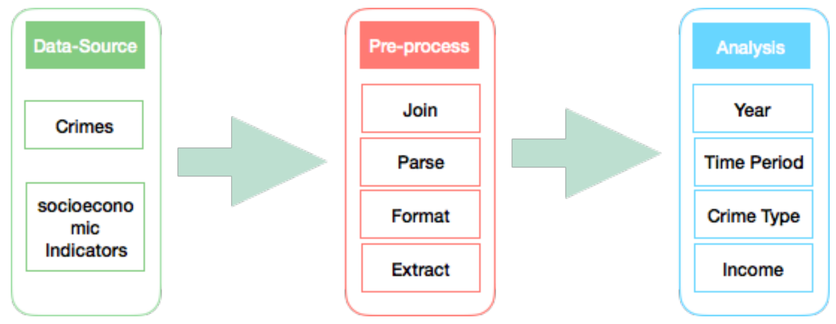
Design and Implementation

To address the four problems described above, there are two main steps: 1) pre-process and 2) analysis as shown in the graph below.

1) Pre-process

In this phase, we join the two dataset together using the community Number attribute. At the same time, we extract the most import attributes we need and throw the rest of attributes away. We also need to parse the date and Per Capital Income attributes and convert them into the format we need in the final analysis.

- Join Social economic factors table with Crime Data using community area Number attribute.
- Parse the Date attribute and extract month, day, and time attributes and broke up a day into 6 time slots(1: 12am – 4am 2: 4am – 8am 3: 8am – 12am 4: 12pm – 4pm 5: 4pm – 8pm 6:8pm – 12pm);
- Scale the income into five slots:[0-14,000), [14,000-18,000), [18,000-24,000), [24,000-35,000), [35,000-).



2) Analysis

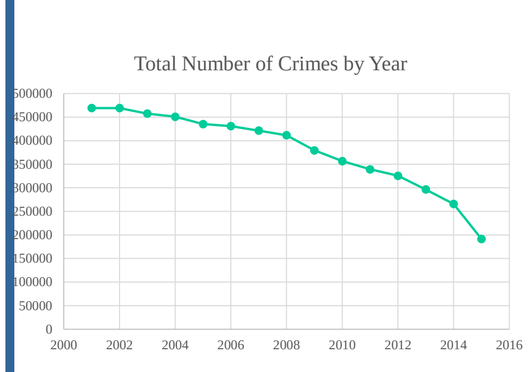
After the phase of pre-processing, we get the intermediate dataset that will be used in the analysis phase which looks like the records below:

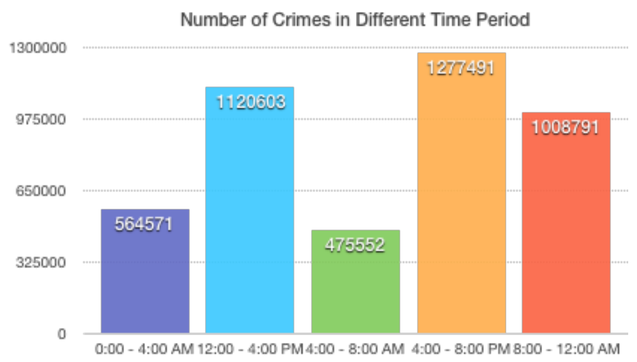
Year	Time Period	Crime Type	Community NO	Community Name	Income
2008	0:00 - 4:00 AM	BATTERY	9	Edison Park	[35000~)

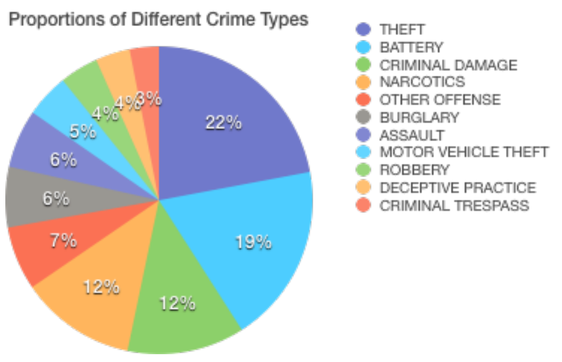
In the phase of Mapper, We chose the Year, Time Period, Crime Type, Income as the key and the number of crimes as the value. In the phase of Reduce, using summarization design pattern, we accumulate the number of crimes respectively.

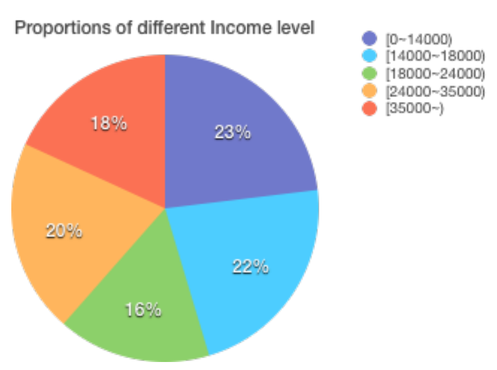
Result

Using visualization tools we visualized the result of Mapper-Reduce as shown below:









Conclusion

By Using Map-Reduce one of the key techniques we learned from the Big Data course, we pre-processed the crime dataset of the city of Chicago and came up with the intermediate result we need as the input of the final analysis. In the phase of analysis, we draw the relationships between the occurrence of crimes and the four main factors. Finally, we visualized the results. The result can be used to understand and guide the real life. However, in fact, the relationships among all of the factors are far more complex than we did here. In the future, we'll consider more factors to analysis and predict crime.